Article

Is routine hospital episode data sufficient for identifying individuals with chronic kidney disease?

Robertson, Lynn M., Denadai, Lucas, Black, Corri, Fluck, Nicholas, Prescott, Gordon, Simpson, William, Wilde, Katie and Marks, Angharad

Robertson, Lynn M., Denadai, Lucas, Black, Corri, Fluck, Nicholas, Prescott, Gordon ORCID: 0000-0002-9156-2361, Simpson, William, Wilde, Katie and Marks, Angharad (2016) Is routine hospital episode data sufficient for identifying individuals with chronic kidney disease? Health Informatics Journal, 22 (2). pp. 383-396. ISSN 1460-4582

Is routine hospital episode data sufficient for identifying individuals with chronic kidney disease? A comparison study with laboratory data

L Robertson[1]
L Denadai[1]
C Black[1,2]
N Fluck[2]
G Prescott[1]
W Simpson[2]
K Wilde[1]
A Marks[1,2]

[1] Division of Applied Health Sciences, University of Aberdeen, Aberdeen, Scotland

[2] NHS Grampian, Aberdeen, Scotland

Corresponding author:
Ms Lynn M Robertson
Division of Applied Health Sciences, University of Aberdeen, Polwarth Building, Foresterhill, Aberdeen, Scotland AB25 2ZD
Email: l.robertson@abdn.ac.uk
Tel: 01224 437135

L Denadai: lucasdenadai89@hotmail.com
C Black: corri.black@abdn.ac.uk
N Fluck: nfluck@nhs.net
G Prescott: gordon.prescott@aben.ac.uk
W Simpson: bill.simpson@nhs.net
K Wilde: k.wilde@abdn.ac.uk
A Marks: a.marks@abdn.ac.uk

ABSTRACT

Internationally, investment in the availability of routine healthcare data for improving health, health surveillance and healthcare is increasing. We assessed the validity of hospital episode data for identifying individuals with chronic kidney disease (CKD) compared to biochemistry data in a large population-based cohort, GLOMMS-II (n=70,435). GLOMMS-II links hospital episode data to biochemistry data for all adults in a health region with impaired kidney function and random samples of individuals with normal and unmeasured kidney function in 2003. We compared identification of individuals with CKD by hospital episode data (based on ICD-10 codes) to the reference standard of biochemistry data (at least two estimated glomerular filtration rates <60 ml/min/1.73m2 at least 90 days apart). Hospital episode data, compared to biochemistry data, identified a lower prevalence of CKD, had low sensitivity (<10%) but high specificity (>97%). Using routine health care data from multiple sources offers the best opportunity to identify individuals with CKD.

INTRODUCTION

Chronic kidney disease (CKD) has been identified as a worldwide public health problem with a rising incidence and prevalence[1], and is associated with high morbidity (cardiovascular disease, need for renal replacement therapy (RRT)), mortality and health care costs (estimated for England 2009-10 to be £1.45 billion[2]). Risk factors for CKD include diabetes, vascular disease, hereditary renal diseases, smoking and hypertension. In 2002, the Kidney Disease Outcomes Quality Initiative (KDOQI) defined and classified CKD based on kidney damage (structural or functional abnormalities of the kidney) with glomerular filtration rate (GFR, a measure of kidney function) ≥60 ml/min/1.73m^2 (stage 1-2) or GFR <60 ml/min/1.73m^2 alone (stage 3-5), present for at least three months[1]. Estimates of prevalence, based on the first part of this definition, in the US suggest the prevalence of CKD stages 1-4

increased from 10.0% in 1988-1994 to 13.1% in 1999-2004.[3] However, other studies have reported varied prevalence rates of CKD (0.6% to 42.6).[4] In UK general practices only 2.9% are registered as having CKD.[5] Part of the variation in prevalence estimates may be due to how CKD is defined and the data sources used to identify individuals with CKD.

For many conditions, information on disease prevalence is estimated from disease registries, GP registers and/or coding of hospital episodes. The use of hospital episode data (recorded in Scotland as the Scottish Morbidity Record (SMR) 01), either as single episodes or longitudinally linked episodes to identify comorbidities has been used extensively in research.[6] For acute events that almost exclusively require hospital admission (e.g. hip fracture) this may be a valid source of information.[7]. For chronic diseases such as CKD, hospital episode data may require supplementation from other sources of data to fully elucidate disease load, and facilitate early identification. The UK government and others internationally, have invested in routine health care data (ie funding opportunities, investment in digital health systems) since it is thought to be important for health and health care through research, health surveillance and health care planning.[8-12]

For individuals with CKD, early detection and management is believed to be important to reduce morbidity and slow progression to RRT.[13] However, the forum of care may vary, with all patients requiring GP care and more advanced patients potentially requiring assessment by nephrology care. In the UK, there is no standard surveillance system for the identification of people with CKD. Ideally those with CKD would be identified clinically from a combination of sources including biochemistry testing for estimated GFR (eGFR) and albuminuria, however this relies on clinicians identifying and noting abnormal results and that these are

sustained abnormalities rather than an acute change. This is sometimes difficult to achieve in regions where biochemistry testing is done by multiple providers and where not all results are returned to a single clinician responsible for compiling results. An alternative means of identifying those with CKD would be to flag those that have routine hospital episode data consistent with this CKD diagnosis and subsequently informing GPs for follow-up and confirmation.

Two recent systematic reviews[14, 15], and recent studies[16-19], have evaluated the degree to which administrative coding accurately identified individuals with kidney diseases, reporting a large variation in sensitivity (3%-88%). Only a few studies have compared hospital administrative data to laboratory data employing the 2002 KDOQI definition of stages 3-5 CKD, of at least two eGFR <60 ml/min/1.73m^2 at least 90 days apart.[18, 20, 21] Of these, only Ronksley *et al*.[18] did so in a community cohort, in Canada. Using a community based population increases the generalisability of results as opposed to relying on, for example, a selected inpatient population. We did not identify any studies from the UK that compared hospital episode data to laboratory data.

With the growing emphasis on the use of routine administrative data, validation studies become increasingly important in order to provide information on the accuracy and validity of findings that are based exclusively on these data. As administrative data have the potential to be a rich source of data for population-based research in CKD, we aimed to assess the validity of diagnostic algorithms for CKD in hospital episode data compared to biochemistry data in a large population-based cohort in Grampian, Scotland.

METHODS

We carried out a validation study within an existing cohort developed by data linkage of biochemistry, hospital episode and death registry data.

Study Population – Grampian Laboratory Outcomes, Morbidity and Mortality Study-II (GLOMMS-II) cohort

All inpatient, out-patient and community serum creatinine (isotope dilution mass spectrometry (IDMS) aligned) and urinary protein measurements in the Grampian region, served by a single United Kingdom National External Quality Assessment Service monitored biochemistry service, are contained in the Grampian Laboratory Renal Database for 1999 to 2009. This database was queried to identify the GLOMMS-II cohort, which was comprised of: all adults (>15 years) with impaired kidney function in 2003; a random sample of individuals with normal or no measure of kidney function in 2003 (but prior and post 2003 sampling); all those with proteinuria but normal kidney function in 2003; and all individuals on renal replacement therapy (RRT) at 1 January 2003 (identified from Scottish Renal Registry and local renal system). Where present, the first "low" eGFR <60 ml/min/1.73m^2 in 2003 was taken as the index value and date. Where all values in 2003 were normal the last value and date were taken as the index. Where no samples were taken in 2003, the index date was taken as 31 December 2003 to allow the potential for the individual to be sampled.

Defining CKD from biochemistry data

eGFR was calculated using the 4 variable IDMS aligned Modification of Diet in Renal Disease (MDRD) formula (serum creatinine, age, sex and race). CKD was defined and staged

according to KDOQI.[1] CKD stage 3-5 (including 3a and 3b) was defined as an index eGFR<60ml/min/1.73m^2 in 2003 followed after 90 days by another low eGFR (<60ml/min/1.73m^2), or if there were no further eGFR values after 90 days post-index, the last eGFR prior to 90 days pre-index also being low i.e. between the start of the database records in 1999 and the index value. CKD stages 1 and 2 were defined as an index eGFR >60ml/min/1.73m^2 with microalbuminuria or macroalbuminuria on urine albumin or protein creatinine ratio (ACR or PCR) testing. Individuals were categorised as not having CKD if their index eGFR was not measured, was normal or was impaired but not CKD (at least one eGFR <60ml/min/1.73m^2 but with no evidence that this was sustained for at least three months).

Defining CKD from hospital episode data

In the UK, information about an episode of hospital care is recorded following a patient's discharge. In Scotland, this information is recorded in the SMR01, which is collated nationally by the Information Services Division (ISD), part of NHS National Services Scotland. SMR01 is an episode-based patient record relating to all inpatient and day case discharges. This information contributes to NHSScotland's Performance Assessment Framework, clinical governance and performance indicators, and for planning and research purposes.[22] Diagnoses are coded using International Classification of Disease-10 (ICD-10) and procedures coded using the Office of Population Censuses and Surveys Classification of Interventions and Procedures (OPCS). We defined CKD for each patient from hospital episode data for two time periods; 2003 (including admission at index) and also adding a five year "look-back" period.

To identify potentially relevant codes to define CKD, an experienced nephrologist reviewed all ICD-10 and OPCS codes. Three groups of codes (algorithms) were developed (Table 1): first, a broad definition encompassing most diseases which might include renal complications ("All codes"); second, an algorithm to define renal disease based on a Charlson comorbidity algorithm[23] ("renal disease"); and third, an algorithm highly likely to identify CKD ("chronic kidney disease").

Table 1: Renal disease-related ICD-10 and OPCS codes (algorithms)

ICD-10/OPCS code	Definition	Coding algorithm definition		
		All codes	Renal Disease	Chronic Kidney Disease
E10.2	Diabetes type 1 with renal complications	•		•
E11.2	Diabetes type 2 with renal complications	•		•
E14.2	Diabetes with renal complications	•		•
I12.0	Hypertensive renal disease with renal failure	•	•	•
I13.1	Hypertensive heart and renal disease with renal failure	•	•	•
M02 (OPCS)	Nephrectomy	•		
N00 to N08	Glomerular diseases	•		
N03.2-N03.7	Chronic nephritic syndrome: - diffuse glomerulonephritis or dense deposit disease	•	•	
N05.2-N05.7	Unspecified nephritic syndrome: - diffuse glomerulonephritis or dense deposit disease	•	•	
N11	Chronic tubule-interstitial nephritis	•		
N13	Obstructive and reflux uropathy	•		
N13.7	Vesicoureteral-reflux-associated uropathy	•		
N18.x	Chronic renal failure	•	•	•
N19.x	Unspecified renal failure	•	•	
N20	Calculus of kidney and ureter (includes nephrolithiasis)	•		
N21	Calculus of lower urinary tract	•		
N22	Calculus of urinary tract in diseases classified elsewhere	•		
N23	Unspecified renal colic	•		
N25.0	Renal osteodystrophy	•	•	
N26	Unspecified contracted kidney	•		
N27	Small Kidney of unknown cause	•		
N28	Ischaemia and infarction of the kidney	•		
Q60	Renal agenesis and other reduction defects of the kidney	•		
Q61	Cystic kidney disease	•		
Q62	Congenital obstructive defects of the renal pelvis and congenital malformation of ureter	•		
Q63	Other congenital malformations of the kidney	•		
Q64	Other congenital malformations of urinary system	•		
Z49.0-Z49.2	Care involving dialysis	•	•	
Z90.5	Acquired absence of kidney	•		
Z94.0	Kidney transplant status	•	•	
Z99.2	Dependence on renal dialysis	•	•	

ICD, International Classification of Diseases; OPCS, Office of Population Censuses and Surveys Classification of Interventions and Procedures

Data linkage

The Community Health Index (CHI) number, a unique patient identifier used throughout the Scottish health care system, was used to link GLOMMS-II with hospital episode data using deterministic matching. Patient identifiers were removed after data linkage. The dataset was stored in the Grampian Data Safe Haven allowing secure controlled access for researchers while ensuring data security.[24]

The flow diagram for generating GLOMMS-II is shown in Figure 1. From the database query 71,251 individuals were identified. There were 471 excluded from the analysis because of missing information on index date, duplication or death on index date. The 345 people already on RRT at index (thus end stage renal disease, not just CKD), were excluded from the analysis (74.8% had a "CKD" code from SMR01). Overall, 70,435 individuals were included in this study.

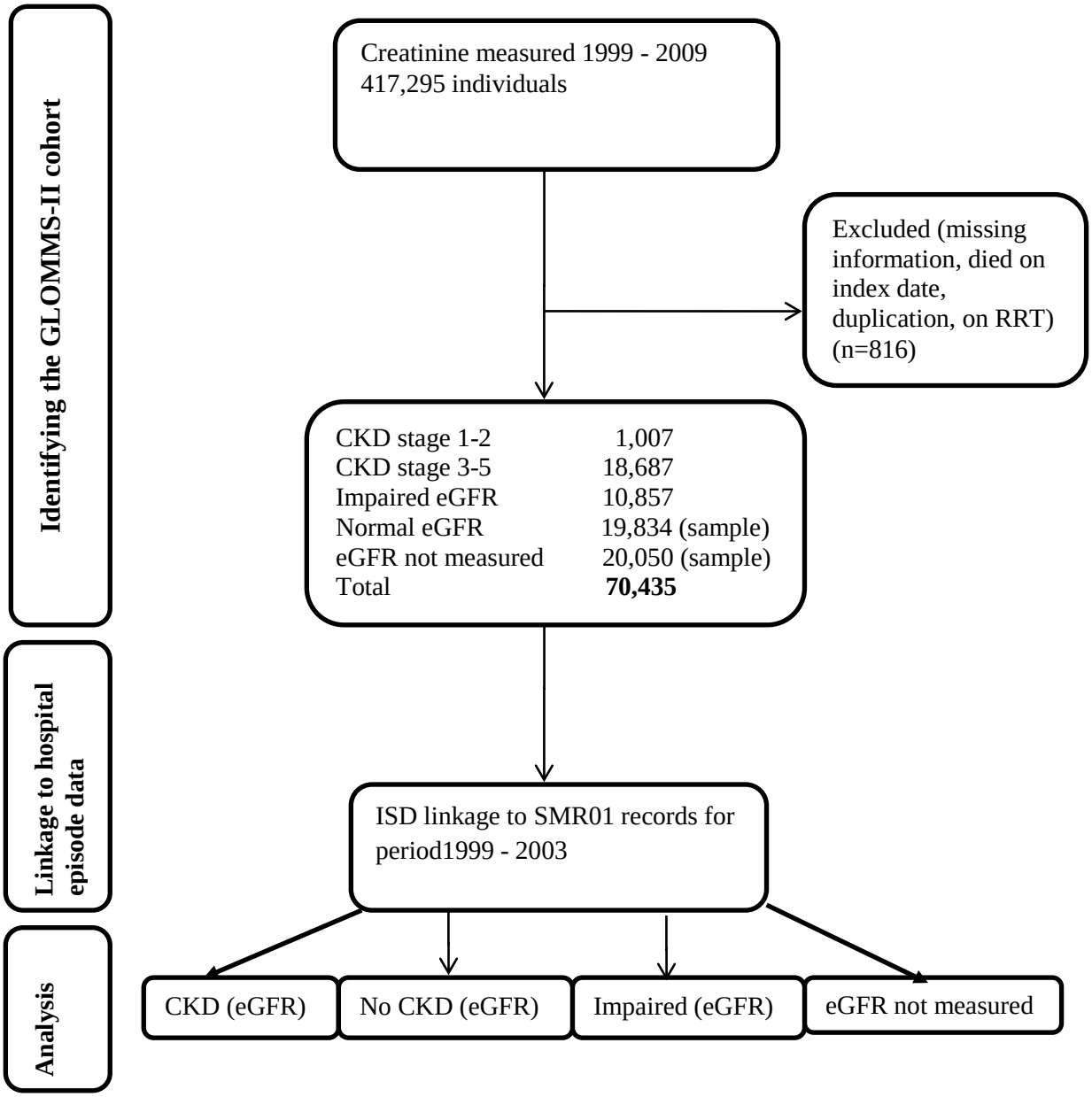

Figure 1: GLOMMS-II flow diagram

Statistical Analysis

Descriptive statistics were used to describe demographic, proteinuria/albuminuria status, creatinine, eGFR and comorbidity variables stratified by renal risk group (CKD stage 1-5/normal eGFR, impaired eGFR or eGFR not measured). Comorbidity was based on the Charlson comorbidity index[25], which is a weighted index that takes into account the number and seriousness of comorbid disease. The proportion of the cohort with CKD identified by biochemistry data and the proportion of the cohort with CKD identified by hospital episode data were calculated. The validity of hospital episode data identified CKD was assessed for the three coding algorithms and two time periods; 2003 (including admission at index) and also adding a five year "look-back" period.

Sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) were calculated against the reference standard of CKD (biochemistry data). Kappa values, κ (a measure of agreement between two sets of categorical measurements on the same individuals)[26], were calculated. We categorised agreement as poor if $\kappa \leq 0.20$, fair if $0.21 \leq \kappa \leq 0.40$, moderate if $0.41 \leq \kappa \leq 0.60$, substantial if $0.61 \leq \kappa \leq 0.80$ and good if $\kappa > 0.80$.[27]

The validity of hospital episode data-defined CKD within specific subgroups was considered, including CKD stage (stage 1-2, 3a, 3b, 4 and 5) and age (<75 or ≥75 years). To explore sensitivity further, analyses were repeated comparing hospital episode data to an alternative definition for biochemistry-defined CKD, which excluded those with impaired eGFR and those with eGFR not measured from the no-CKD definition. Analyses were performed using Stata version 13[28] and Microsoft Excel.

RESULTS

A total of 70,435 individuals were included. The characteristics of the study population are shown in Table 2. Based on biochemistry data, 28% (19,694) of the cohort had CKD stage 1-5 (which equates to 4.5% of the adult Grampian population in 2003 (433,109)[29]). Overall, the median age of the cohort was 63.3 years and 58.4% were female. As expected, those with CKD were older than those with normal eGFR or "not measured". Charlson comorbidity categories for CKD Stage 1-5 and impaired eGFR were similarly distributed with more than two-thirds of individuals with a score of zero. Those with normal eGFR or "not measured" in 2003 had the lowest Charlson scores. Of note, there were 63 individuals with macroalbuminuria but no eGFR measured. Of those with CKD identified by biochemistry, 6,767 individuals had no hospital admission in the five years prior to 2003.

As shown in Table 3, based on the reference standard of biochemistry-defined CKD, 28% (19,694) of the cohort had CKD stage 1-5. The proportion of the cohort identified with probable CKD by hospital episode data was substantially lower, ranging from 0.8% to 4.1% over the three coding algorithms and two time periods.

Hospital episode data identified CKD was generally less common compared to biochemistry-defined CKD, and varied across coding algorithms and time periods (Table 3). The sensitivity of hospital episode coding compared to biochemistry for identifying CKD was low, ranging from 2.2% to 8.6%. Specificity of coding was >97% for all coding algorithms and time periods. All algorithms improved by adding a five year look-back period in addition to just SMR01 records from 2003, showing higher sensitivities. The very inclusive “all codes”

algorithm was most sensitive but least specific, followed by the "renal disease" and "chronic kidney disease" algorithm which was most specific. Overall the agreement between hospital episode data and biochemistry defined CKD was very poor (kappa values <0.1) because of low numbers identified with hospital episode data, despite excellent specificity.

Sensitivity analyses were carried out comparing hospital episode data to an alternative definition for biochemistry-defined CKD, excluding those with impaired eGFR and those with eGFR not measured from the no-CKD definition. However, this, as expected, only improved the PPV further and reduced the NPV further of hospital episode data. For those with CKD algorithm defined CKD using 2003 plus five year look-back data, PPV 99.56% (vs 81.06%) and NPV 51.05% (vs 72.68%).

Using the "renal disease" and "CKD" coding algorithms, since more specific, including the five-year look-back period, the performance within age and CKD stage subgroups was considered (Table 4). Amongst those with biochemistry identified CKD, the "renal disease" algorithm identified similar but slightly more individuals than the "CKD" algorithm. Worse CKD stage was associated with better identification (sensitivity) using both hospital episode based algorithms (4.8% of stage 3b compared to 56.9% of stage 5 CKD, for the "CKD" algorithm). For biochemistry identified CKD stage 3b to 5, younger age (<75 vs ≥75 years) was associated with a higher sensitivity using the hospital episode recording algorithms.

Table 2: Characteristics of study population

Characteristic	Total		Chronic kidney disease				Not chronic kidney disease					
			CKD stage3-5		CKD stage1-2		Impaired		Normal eGFR		eGFR not measured	
	n	(%)	*n*	(%)	*n*	(%)	*n*	(%)	*n*	(%)	*n*	(%)
Total	70435	100.0	18687	100.0	1007	100.0	10857	100.0	19834	100.0	20050	100.0
Sex												
Male	29322	(41.6)	6580	(35.2)	649	(64.5)	4323	(39.8)	9346	(47.1)	8424	(42.0)
Female	41113	(58.4)	12107	(64.8)	358	(35.6)	6534	(60.2)	10488	(52.9)	11626	(58.0)
Age at index (years), Median (IQR)	63.3	(47.8- 75.5)	75.7	(68.5- 82.0)	61.0	(49.3- 69.8)	71.0	(60.7- 79.6)	53.1	(38.9- 65.5)	52.0	(39.0- 64.8)
15-44 years	15259	(21.7)	305	(1.6)	195	(19.4)	643	(5.9)	6909	(34.8)	7207	(36.0)
45-54 years	9690	(13.8)	660	(3.5)	169	(16.8)	980	(9.0)	3802	(19.2)	4079	(20.3)
55-64 years	12375	(17.6)	2201	(11.8)	259	(25.7)	2147	(19.8)	3966	(20.0)	3082	(19.0)
65-74 years	14788	(21.0)	5630	(30.1)	249	(24.7)	2945	(27.1)	3207	(16.2)	2757	(13.8)
75-84 years	13404	(19.0)	7119	(38.1)	123	(12.2)	2825	(26.0)	1624	(8.2)	1713	(8.5)
≥85 years	4919	(7.0)	2772	(14.8)	12	(1.2)	1317	(12.1)	326	(1.6)	492	(2.5)
PCR at index(n=1845), median (IQR)	22	(10- 62)	27	(11- 72)	114	(73- 212)	16	(9- 38)	10	(6- 18)	17	(8- 38)
ACR at index(n=5439), median (IQR)	1	(1- 6)	1	(1- 6)	8	(5- 17)	1	(1- 3)	1	(1- 1)	4	(1- 10)
Proteinuria status												
Untested	63158	(89.7)	15412	(82.5)	0	(0.0)	9593	(88.4)	18602	(93.8)	19551	(97.5)
Normoalbuminuric	4580	(6.5)	2125	(11.4)	0	(0.0)	942	(8.7)	1232	(6.2)	281	(1.4)
Microalbuminuric	1725	(2.5)	602	(3.2)	768	(76.3)	200	(1.8)	0	(0.0)	155	(0.8)
Macroalbuminuric	972	(1.4)	548	(2.9)	239	(23.7)	122	(1.1)	0	(0.0)	63	(0.3)
Creatinine at index, median (IQR)	85.5	(71.4- 103.8)	108.1	(91.9- 126.4)	79.0	(68.2- 87.6)	102.7	(87.6- 115.7)	73.6	(65-.0 84.4)	74.7	(65-.0 85.5)
eGFR (ml/min/1.73m²) at index, median (IQR)	66.8	(53.5- 85.2)	49.7	(41.6- 55.2)	79.8	(71.2- 91.2)	55.3	(49.9- 58.1)	82.2	(72.7- 94.3)	82.3	(71.4- 95.2)
Charlson comorbidity index group												
0	56242	(79.9)	12667	(67.8)	671	(66.6)	7190	(66.2)	17074	(86.1)	18640	(93.0)
1-2	11308	(16.1)	4763	(25.5)	275	(27.3)	2693	(24.8)	2281	(11.5)	1296	(6.5)
3-4	1943	(2.8)	946	(5.1)	45	(4.5)	598	(5.5)	285	(1.4)	69	(0.3)
≥5	942	(1.3)	311	(1.7)	16	(1.6)	376	(3.5)	194	(1.0)	45	(0.2)

CKD, chronic kidney disease; eGFR, glomerular filtration rate; IQR, interquartile range; PCR, protein-creatinine ratio; ACR, albumin-creatinine ratio
Renal risk groups based on biochemistry data

Table 3: Validity of hospital episode data definition for chronic kidney disease compared to the reference standard of biochemistry

Algorithm/time period	Biochemistry+ HE Coding+	Biochemistry+ HE Coding-	Biochemistry- HE Coding-	Biochemistry- HE Coding+	Proportion of cohort identified with CKD				Hospital episode coding				
	True positive	False negative	True negative	False positive	Biochemistry		HE Coding		PPV	NPV	Sensitivity	Specificity	Kappa*
					n	%	*n*	%					
					19694	28.0%							
All codes													
2003	840	18854	50249	492			1332	1.9%	63.06%	72.72%	4.27%	99.03%	0.0461
2003 plus 5 year lookback	1689	18005	49508	1233			2922	4.1%	57.80%	73.33%	8.58%	97.57%	0.0831
Renal Disease													
2003	595	19099	50527	214			809	1.1%	73.55%	72.57%	3.02%	99.58%	0.0368
2003 plus 5 year lookback	1022	18672	50378	363			1385	2.0%	73.79%	72.96%	5.19%	99.28%	0.0625
Chronic Kidney Disease													
2003	441	19253	50643	98			539	0.8%	81.82%	72.45%	2.24%	99.81%	0.0291
2003 plus 5 year lookback	676	19018	50583	158			834	1.2%	81.06%	72.68%	3.43%	99.69%	0.0441

HE, hospital episode; CKD, chronic kidney disease; PPV, positive predictive value; NPV, negative predictive value

*Interpretation of kappa: Agreement poor if $\kappa \leq 0.20$, fair if $0.21 \leq \kappa \leq 0.40$, moderate if $0.41 \leq \kappa \leq 0.60$, substantial if $0.61 \leq \kappa \leq 0.80$ and good if $\kappa > 0.80$.

Biochemistry definition of CKD/no CKD: Stage 1-5/normal, impaired or not measured (see Methods section).

Hospital episode coding (SMR01) definition of CKD/no CKD: ICD and OPCS codes as detailed in Table 1/no coding or no admission (see Methods section).

Table 4: Validity of hospital episode coding definition (2003 +5 year look-back) for chronic kidney disease compared to the reference standard of biochemistry by stage and age group

	Biochemistry+					
	HE Coding+	**HE Coding-**	**PPV**	**NPV**	**Sensitivity**	**Specificity**
Renal Disease						
Stage 5	88	56	19.5%	99.9%	61.1%	99.3%
Age <75 years	53	29	25.4%	99.9%	64.6%	99.6%
Age ≥75 years	35	27	14.5%	99.7%	56.5%	97.5%
Stage 4	330	916	47.6%	98.2%	26.5%	99.3%
Age <75 years	137	254	46.8%	99.4%	35.0%	99.6%
Age ≥75 years	193	662	48.3%	92.4%	22.6%	97.5%
Stage 3b	388	4563	51.7%	91.7%	7.8%	99.3%
Age <75 years	176	1553	53.0%	96.5%	10.2%	99.6%
Age ≥75 years	212	3010	50.6%	72.9%	6.6%	97.5%
Stage 3a	207	12139	36.3%	80.6%	1.7%	99.3%
Age <75 years	107	6487	40.7%	86.7%	1.6%	99.6%
Age ≥75 years	100	5652	32.6%	58.9%	1.7%	97.5%
Stage 1 and 2	9	998	2.4%	98.1%	0.9%	99.3%
Age <75 years	7	865	4.3%	98.0%	0.8%	99.6%
Age ≥75 years	<5	133	1.0%	98.4%	1.5%	97.5%
Chronic Kidney Disease						
Stage 5	82	62	34.2%	99.9%	56.9%	99.7%
Age <75 years	53	29	50.0%	99.9%	64.6%	99.9%
Age ≥75 years	29	33	21.6%	99.6%	46.8%	98.7%
Stage 4	254	992	61.7%	98.1%	20.4%	99.7%
Age <75 years	111	280	67.7%	99.3%	28.4%	99.9%
Age ≥75 years	143	712	57.7%	92.0%	16.7%	98.7%
Stage 3b	239	4712	60.2%	91.5%	4.8%	99.7%
Age <75 years	118	1611	69.0%	96.3%	6.8%	99.9%
Age ≥75 years	121	3101	53.5%	72.5%	3.8%	98.7%
Stage 3a	100	12246	38.8%	80.5%	0.8%	99.7%
Age <75 years	59	6535	52.7%	86.6%	0.9%	99.9%
Age ≥75 years	41	5711	28.1%	58.9%	0.7%	98.7%
Stage 1 and 2	<5	1006	0.6%	98.0%	0.1%	99.7%
Age <75 years	<5	871	1.9%	98.0%	0.1%	99.9%
Age ≥75 years	0	135	0.0%	98.4%	0.0%	98.7%

HE, hospital episode**.**

Biochemistry definition of CKD/no CKD: Stage 1-5/normal, impaired or not measured (see Methods section).
Hospital episode coding (SMR01) definition of CKD/no CKD: ICD and OPCS codes as detailed in Table 1/no coding or no admission (see Methods section).

DISCUSSION

We used a large UK community cohort to demonstrate whether the use of coding algorithms to identify renal disease, in particular CKD, from hospital episode data was a useful alternative should biochemistry data be difficult to access. We found that hospital episode data coding algorithms were very specific for CKD, however sensitivities were very poor (at best only 8.6% identified), as was agreement. Of interest the proportion of those with CKD identified through biochemistry data who were also identified with hospital episode coding was higher at more advanced CKD stages and in those under 75 years of age.

CKD is recorded poorly in hospital episode data. This may be because CKD is often not the main reason for admission. This is likely to be similar for other chronic diseases such as diabetes and hypertension, unlike acute events such as hip fracture. Also, the recognition of CKD in the time prior to eGFR reporting (2008) was poor, and may have improved in the time since then. Those with more advanced renal disease are also more likely to be frequent in-patients as a result of the higher comorbidity load[30] and as a result of increased complications of their renal disease, thus the more likely that renal disease will be recognised during the admission episode coding.

Comparison with existing literature

Few studies[18, 20, 21] have validated hospital administrative data compared with a reference standard of biochemistry data employing the KDOQI definition of CKD, of at least two eGFR <60 ml/min/1.73m^2 at least 90 days apart, and none included CKD stage 1 and 2 (those with proteinuria). In keeping with our findings, where reported, sensitivities are low and specificities high for hospital episode data compared to biochemistry defined CKD.[14, 15, 18, 19]

We also found high PPVs, which means that individuals who are identified as having CKD from hospital episode coding, do have CKD according to biochemistry data, thus any diagnosis based on coding should be accurate using the algorithms outlined, although very un-sensitive. The range of PPV values reported in other CKD validation studies has been broad (29%-100%).[15, 18]

Our study used a very large population-based cohort. Only one other study has used a community based population.[18] However, Ronksley *et al*. looked for hospital episode data after the biochemistry identification of CKD. Therefore they were assessing whether those with CKD were being identified at their next hospital admission, not whether a prevalence cohort with CKD was identifiable equally from biochemistry or hospital episode coding.[18] This use of a three year window after biochemistry identified disease would perhaps identify patients too late for intervention, thus our method is perhaps more applicable for identifying those with disease.

We have demonstrated that those with more advanced CKD are more likely to be captured by hospital episode data, also reported by others.[18, 21] This is in keeping with the fact that at the time of this study, eGFR reporting had not been instigated in the UK and as such, the identification of CKD would be expected only in those with more advanced CKD, both by clinicians and SMR01 coders. Ronksley *et al*. reported that estimates of sensitivity were higher when eGFR <30 mL/min/1.73m^2 was used as the reference standard compared with using <60 mL/min/1.73m^2.[18] Ferris *et al*. reported a similar pattern in in-patients.[21]

Studies have reported that older age was not significantly associated with a greater likelihood of being labelled with CKD.[21] However, this was a study of inpatients, therefore the risk profile identified with biochemistry might have been different. Our finding that younger individuals with CKD were identified better on hospital episode data than older individuals has been previously reported.[18] For younger individuals, CKD is likely to be more of a significant problem than for those that are elderly with CKD with the same degree of renal impairment. It may also reflect that those with CKD at younger ages are likely to have fewer comorbidites when admitted to hospital and therefore have this recognised when discharge coding is carried out.[31]

Denburg *et al*.[17] looked at the recording of biochemistry results at a general practice level compared to the recognition of CKD on general practice coding, which again found low sensitivity but excellent specificity and high PPV. It is unclear, however, how many of the biochemistry results had been entered into GP systems manually.

Strengths and limitations

This study has many strengths. It is one of only a few studies assessing agreement between biochemistry defined CKD that was required to be present for greater than three months compared to hospital episode data.[18, 20, 21] It is a very large population-based cohort, not limited to a specific patient group, and since ICD-10 coding is used, we might expect these findings to be potentially generalisable to other chronic diseases, eg diabetes, and across the world. The universal nature of the biochemistry service to the region ensures that those living within the region who have testing of renal function would have results available for

consideration, and where repeated these would be available, assisting in the identification of those with truly *chronic* kidney disease.

There are, however, limitations to this study. Calculating eGFR using the MDRD equation is reflective of current UK practice and thus the individuals currently identified as having CKD, however there are others outside of the UK who support the use of the CKD-EPI equation. It would be expected that both eGFR equations would identify similar individuals with CKD, particularly at more advanced stages, and it is unlikely that the results would be significantly different.[31] The use of only hospital episode data as a source of confirmatory CKD recording, although fulfilling the aim of this paper to ascertain its validity, meant that other routine sources of such data such as GP coding, were not assessed. Although this would be a useful additional source of data, it was not available to us, would require assessment in its own right, and has been explored at least at a GP biochemistry recording level before.[17] Our biochemistry definition of no-CKD was all-inclusive, including impaired eGFR (at least one eGFR <60ml/min/1.73m^2 but not sustained) and eGFR not measured. However, we performed sensitivity analyses, defining "no CKD" as those with normal eGFR only and found that this only improved PPV and worsened NPV. Sensitivity and specificity were similar. As noted previously, the recognition of CKD in the time prior to eGFR reporting (2008) was poor, and may have improved in the since then. However, this is unlikely to change the greater sensitivity of eGFR reporting over SMR01.

Implications for future research or clinical practice

As mentioned in the introduction, hospital episode data may be sufficient for acute hospital care requiring events. However, for chronic conditions, as illustrated here with CKD, the use

of corroborating additional data when admissions are due to another event or comorbidity may be necessary.

As demonstrated, hospital episode coding data is very specific with high PPV for the identification of individuals with CKD. This has implications for both clinical practice and future research. With clinical practice, it is insufficient to use hospital episode data alone to identify those with CKD, and access to current and historical biochemistry data is essential to identifying CKD appropriately. However, the use of hospital episode data as an additional flag is potentially useful for identifying high risk individuals. Another issue for clinical practice is patient safety, particularly with the prescribing of drugs that are either nephrotoxic or with significant renal clearance. The use of both systems of identification should improve patient safety issues related to this. This also applies to preparation for surgical, radiological and oncological procedures.

For research, we have demonstrated that biochemistry data is crucial for describing the prevalence of CKD and therefore the healthcare burden associated with it, not just the few identified through hospital episode data. Historically, CKD identified through hospital episode coding described high RRT initiation rates. However, in cohorts identified through biochemistry more recently, the rates reported have been lower.[32] Whether this is due to the severity of CKD identified being different, or due to the disease processes being different, is not clear and requires further research. There are also implications for clinical trials, in that the event rate that sample sizes are based on may differ depending on the source of CKD identification.

The ideal for the future would be a unifying electronic patient healthcare record containing information on previous hospital identified events, general practice and also biochemistry results, to ensure accurate and timely identification of those with CKD.

Conclusion

The findings of this study suggest that routine hospital episode data has limited value in the routine identification of individuals with CKD. However where those with CKD have been identified using hospital episode data, this information is highly specific. Other sources of routine health care data such as routine biochemistry data, including historical data, and not just that pertaining to a given event, should be available to clinicians caring for patients, and are an important source for further research into clinical outcomes, including hospitalisations. The most important uses of this data are for planning, surveillance, screening, and for research.

Ethics

The study protocol was reviewed by the Privacy Advisory Committee for ISD, NHS Grampian Caldicott Guardian. The North of Scotland Research Ethics Service reviewed the project and felt it was audit rather than research. The College Ethics Review Board of the University of Aberdeen, College of Life Sciences and Medicine also reviewed the protocol. There were no concerns.

Acknowledgments

We thank Information Services Division Scotland who provided the SMR01 data and NHS Grampian who provided the biochemistry data. We also thank the University of Aberdeen Data Management Team.

Declaration of Conflicting Interests

None

Funding

This work was supported by the Chief Scientists Office for Scotland [grant number CZH/4/656].

REFERENCES

1. National Kidney Foundation. K/DOQI clinical practice guidelines for chronic kidney disease: evaluation, classification, and stratification. Am J Kidney Dis 2002; 39 Suppl 1: S1-S266.

2. Kerr M, Bray B, Medcalf J, et al. Estimating the financial cost of chronic kidney disease to the NHS in England. Nephrol Dial Transplant 2012; 27 Suppl 3: iii73-80.

3. Coresh J, Selvin E, Stevens LA, et al. Prevalence of chronic kidney disease in the United States. JAMA 2007; 298: 2038-2047.

4. McCullough K, Sharma P, Ali T, et al. Measuring the population burden of chronic kidney disease: a systematic literature review of the estimated prevalence of impaired kidney function. Nephrol Dial Transplant 2012; 27: 1812-1821.

5. Walker N, Bankart J, Brunskill N, et al. Which factors are associated with higher rates of chronic kidney disease recording in primary care? A cross-sectional survey of GP practices. Br J Gen Pract 2011; 61: 203-205.

6. Leal JR and Laupland KB. Validity of ascertainment of co-morbid illness using administrative databases: a systematic review. Clin Microbiol Infect 2010; 16: 715-721.

7. Hudson M, Avina-Zubieta A, Lacaille D, et al. The validity of administrative data to identify hip fractures is high--a systematic review. J Clin Epidemiol 2013; 66: 278-285.

8. Department of Health, NHS Improvement & Efficiency Directorate,Innovation and Service Improvement. Innovation Health and Wealth, Accelerating Adoption and Diffusion in the NHS.
http://webarchive.nationalarchives.gov.uk/20130107105354/http://www.dh.gov.uk/prod_consum_dh/groups/dh_digitalassets/documents/digitalasset/dh_134597.pdf; (2011, accessed September 2014)

9. Medical Research Council. Funding Opportunities, E-Health Informatics Research Centres (E-HIRCs) Call, http://www.mrc.ac.uk/Fundingopportunities/Calls/E-healthCentresCall/index.htm; (2011, accessed May 2013).

10. Medical Research Council. Strategic Framework for Health Informatics in Support of Research, http://www.mrc.ac.uk/Utilities/Documentrecord/index.htm?d=MRC006669 (2010, accessed May 2013).

11. Medical Research Council. UK e-health records research capacity and capability, http://www.mrc.ac.uk/Utilities/Documentrecord/index.htm?d=MRC007896 (2011, accessed May 2013).

12. The White House, Office of Science and Technology Policy. Big data is a big deal, http://www.whitehouse.gov/blog/2012/03/29/big-data-big-deal (2012, accessed September 2014).

13. Black C, Sharma P, Scotland G, et al. Early referral strategies for management of people with markers of renal disease: a systematic review of the evidence of clinical effectiveness, cost-effectiveness and economic analysis. Health Technol Assess 2010; 14: 21.

14. Grams ME, Plantinga LC, Hedgeman E, et al. Validation of CKD and related conditions in existing data sets: A systematic review. Am J Kidney Dis 2011; 57: 44-54.

15. Vlasschaert ME, Bejaimal SA, Hackam DG, et al. Validity of administrative database coding for kidney disease: a systematic review. Am J Kidney Dis 2011; 57: 29-43.

16. Chase HS, Radhakrishnan J, Shirazian S, et al. Under-documentation of chronic kidney disease in the electronic health record in outpatients. J Am Med Inform Assoc 2010; 17: 588-594.

17. Denburg MR, Haynes K, Shults J, et al. Validation of The Health Improvement Network (THIN) database for epidemiologic studies of chronic kidney disease. Pharmacoepidemiology & Drug Saf. 2011; 20: 1138-1149.

18. Ronksley PE, Tonelli M, Quan H, et al. Validating a case definition for chronic kidney disease using administrative data. Nephrol Dial Transplant 2012; 27: 1826-1831.

19. Fleet JL, Dixon SN, Shariff SZ, et al. Detecting chronic kidney disease in population-based administrative databases using an algorithm of hospital encounter and physician claim codes. BMC Neph 2013; 14: 81.

20. Kern EFO, Maney M, Miller DR, et al. Failure of ICD-9-CM codes to identify patients with comorbid chronic kidney disease in diabetes. Health Serv Res 2006; 41: 564-580.

21. Ferris M, Shoham DA, Pierre-Louis M, et al. High prevalence of unlabeled chronic kidney disease among inpatients at a tertiary-care hospital. Am J Med Sci 2009; 337: 93-97.

22. Information and Statistics Division. NHS Scotland Data Quality Assurance Report on Acute Inpatient/Day Case Data 2000 - 2002. Edinburgh:NHSScotland. (http://www.isdscotland.org/Products-and-Services/Data-Quality/Previous-Projects/SMR01%20National%20Report.pdf)

23. Quan H, Sundararajan V, Halfon P, et al. Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data. Med Care 2005; 43: 1130-1139.

24. University of Aberdeen. Grampian Data Safe Haven, http://www.abdn.ac.uk/iahs/facilities/grampian-data-safe-haven.php (2013, accessed September 2014).

25. Charlson M, Pompei P, Ales K, et al. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. J Chron Disease 1987; 40: 373.

26. Landis JR and Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977; 33: 159-174.

27. Petrie A and Sabin C. *Assessing Agreement*. Oxford, England: Oxford, England, 2000, p.93.

28. StataCorp. Stata Statistical Software: Release 13 2013. College Station. TX:StataCorp LP.

29. National Records of Scotland. Revised Mid-2003 Population Estimates, Council and Health Board Areas, http://www.gro-scotland.gov.uk/statistics/theme/population/estimates/mid-year/archive/2003/index.html (2011, accessed October 2013).

30. James MT, Quan H, Tonelli M, et al. CKD and risk of hospitalization and death with pneumonia. Am J Kidney Dis 2009; 54: 24-32.

31. Soo M, Robertson LM, Ali T, et al. Approaches to ascertaining comorbidity information: validation of routine hospital episode data with clinician-based case note review. BMC Res Notes 2014; 7: 253

32. White SL, Polkinghorne KR, Atkins RC, et al. Comparison of the prevalence and mortality risk of CKD in Australia using the CKD Epidemiology Collaboration (CKD-EPI) and Modification of Diet in Renal Disease (MDRD) Study GFR estimating equations: the AusDiab (Australian Diabetes, Obesity and Lifestyle) Study. Am J Kidney Dis 2010; 55: 660-670.

33. Marks A, Black C, Fluck N, et al. Translating chronic kidney disease epidemiology into patient care--the individual/public health risk paradox. Nephrol Dial Transplant 2012; 27 Suppl 3: iii65-72.